What is claimed is:

1. A method of retrieving data, relevant to a topic of interest, from at least one collection of documents, comprising the steps of:

selecting at least one collection;

applying a test query to said at least one collection, thereby retrieving a first set of documents from said at least one collection, said test query including at least one test query term;

classifying each document in a representative sample of documents within said first set of documents according to their relevance to said topic;

extracting all phrases from said first set of documents

selecting high frequency, high technical content phrases from said extracted phrases;

performing a phrase frequency analysis of at least group of said first set of documents having a greater relevance to said subject matter than other documents within said first set of documents to generate a list of phrases including phrase frequency data for each listed phrase;

grouping said selected high frequency, high technical content phrases into thematic categories;

identifying at least one anchor phrase within said phrase frequency analyzed documents for each of said thematic categories;

analyzing phrase co-occurrence of phrases in said phrase frequency analyzed documents to generate a list of co-occurrence pairs, each said co-occurrence pair consisting of an anchor phrase and another listed phrase, to generate a list of co-occurrence pairs including co-occurrence data for each listed co-occurrence pair;

combining said list of phrases with said list of co-occurrence pairs to form a list of candidate query terms;

selecting a plurality of listed query terms from said list of candidates;

applying an additional query to said at least one collection, said additional query being said plurality of said listed query terms, thereby retrieving an additional set of documents from said at least one collection;

classifying at least a representative sample of documents within said additional set of

documents according to their relevance to said topic;

determining, based upon said classification of said representative sample of said additional set of documents, the ratio of relevant to non-relevant documents that are retrieved by each term of said selected plurality of listed query terms;

building a narrowed query consisting of those listed query terms within said plurality of query terms for which said ratio is above a predetermined lower limit;

applying said narrowed query to said collection, thereby retrieving another set of documents from said at least one collection.

2. The method of claim 1, wherein said ratio is equal to:

the number of times a listed query term appears in said set of relevant documents divided by the number of times said listed query term appears in said set of non-relevant documents; or

the number of relevant documents including a listed query divided by the number of non-relevant documents including said listed query term.

3. The method of claim 1, further comprising the step of tagging said at least one anchor phrase and each phrase within all query terms selected for application to said collection during the selection of said plurality of listed terms from said list of candidates, said tagging distinguishing between anchor phrases, selected query terms, and non-query terms subsumed within said selected query terms.

4. The method of claim 1, further comprising the step of choosing one or more documents before selecting said at least one collection, and wherein all documents in said at least one collection cite said chosen one or more documents, or wherein all documents in said at least one collection are cited by said chosen one or more documents.

5. The method of claim 4, wherein said at least one collection consists of all documents citing said chosen one or more references, or wherein said at least one collection consists of all documents cited by said chosen one or more documents.

6. The method of claim 1, wherein at least one of said selected listed query terms is bibliographic.

7. The method of claim 6, wherein said at least one of said selected listed query terms is author name, journal name, or institution name.

8. The method of claim 6, wherein all of said selected listed query terms are bibliographic.

9. The method of claim 8, wherein said at least one of said selected listed query terms is author name, journal name, or institution name.

10. A method of determining levels of emphasis, comprising the steps of:

selecting a collection of documents, each document containing at least one unstructured field;

extracting all phrases from said unstructured field;

filtering all extracted phrases to generate a list of high technical content phrases;

generating a co-occurrence matrix of high technical content phrases for said unstructured field;

normalizing matrix cell values of said co-occurrence matrix to generate a normalized matrix for said field;

grouping phrases from said unstructured field by clustering techniques on said normalized matrix;

summing the phrase frequencies of occurrence within each group, thereby indicating a level of emphasis for each group generated from said collection.

11. The method of claim 10, wherein said normalization is achieved by equivalence index or inclusion index and said normalized matrix for said field is a normalized co-occurrence matrix.

12. The method of claim 10, wherein said normalization is achieved by standard statistical techniques and said normalized matrix for said field is a normalized correlation matrix.

13. The method of claim 12, wherein said filtering step comprises:

generating a factor matrix from said correlation matrix, each phrase being assigned a factor loading value for each factor within said factor matrix, said factor loading value representing the strength of association between said phrase and said factor;

selecting, as said high technical content phrases, phrases having a factor loading value which, for at least one of said factors, is above a threshold value.

14. The method of claim 13, wherein said filtering step further comprises:

determining, for at least one said factor, a difference in the factor loading values between more than one variant of a single phrase;

setting a threshold value for said difference, said threshold value being a value at or below which said variants are deemed to be similar;

establishing, based upon said comparison, whether said variants are similar or dissimilar with respect to said at least one factor;

conflating only said similar variants of said single phrase into a common phrase before selecting said high technical content phrases.

15. The method of claim 13, further comprising the steps of:

determining the difference between the factor loading values of two or more of said phrases in the factor matrix with respect to a single factor in said matrix;

setting a threshold value for said difference, said threshold value being a value at or below which said phrases are deemed to be similar;

establishing, based upon said comparison, whether said two or more phrases in said

factor matrix are similar or dissimilar with respect to said factor.

16. The method of claim 10, wherein said collection of documents includes an additional field, and further comprising the steps of:

extracting all phrases from said additional field;

filtering all extracted phrases to generate a list of high technical content phrases;

generating a co-occurrence matrix of high technical content phrases for said additional field;

normalizing matrix cell values of said co-occurrence matrix to generate a normalized matrix for said additional field;

generating a cross-field co-occurrence matrix of high technical content phrases for said at least one field and said additional field;

normalizing matrix cell values of said cross-field co-occurrence matrix to generate normalized cross-field matrices; and

grouping phrases for each of said unstructured field by clustering techniques on each of said normalized matrices.

17. The method of claim 16, wherein said additional field is structured.

18. The method of claim 16, wherein said additional field is unstructured.

19. The method of claim 10, further comprising the step of choosing one or more documents before selecting said at least one collection, and wherein all documents in said at least one collection cite said chosen one or more documents, or wherein all documents in said at least one collection are cited by said chosen one or more documents.

20. A method of classifying documents retrieved during a collection search, comprising the steps of:

performing a phrase frequency analysis upon said documents to obtain theme and sub-

theme relationships and taxonomies of all high technical content phrases in said documents.

21. A method of generating a taxonomy of a collection of documents, comprising the steps of:

selecting a collection of documents, each document containing at least one structured field;

extracting all phrases from said structured field;

factor matrix filtering all of said extracted phrases to generate a list of high technical content phrases;

generating a co-occurrence matrix of said listed phrases for said field;

normalizing cell values of said co-occurrence matrix to generate a normalized matrix for said field;

grouping said listed phrases for each said field using clustering techniques on said normalized matrix.

summing the frequencies of occurrence within each group, thereby indicating a level of emphasis for each group generated from said collection.

22. The method of claim 21, wherein said collection of documents includes an additional field, and further comprising the steps of:

extracting all phrases from said additional field

factor matrix filtering all of said extracted phrases to generate a list of high technical content phrases;

generating a co-occurrence matrix of said listed phrases for said additional field;

normalizing cell values of said co-occurrence matrices to generate a normalized matrix for said additional field;

generating at least one cross-field co-occurrence matrix for said at least one field and said additional field;

normalizing cross-field co-occurrence matrix cell values to generate a normalized cross-

field matrix;

grouping said text elements for both of said fields using clustering techniques on each of said normalized matrices.

23. The method of claim 22, wherein said additional field is structured.

24. The method of claim 22, wherein said additional field is unstructured.

25. A method of literature-based problem solving, comprising the steps of:

identifying a problem;

selecting a source database comprising documents related to said problem, each of said documents including at least one unstructured field;

retrieving all documents relevant to said problem from said source database to form a set of initially retrieved documents;

extracting all phrases from said unstructured field of said set of initially retrieved documents;

factor matrix filtering all of said extracted phrases to generate a first list of high technical content phrases;

generating a co-occurrence matrix of said high technical content phrases from said first list;

normalizing matrix cell values of said co-occurrence matrix to generate a normalized matrix for said field;

grouping phrases from said unstructured field into thematic categories and subcategories by clustering techniques on said normalized matrix;

generating a directly related topical literature for each said subcategory by retrieving documents related to each of said subcategories, each said directly related topical literature being disjoint with said selected documents and with said directly related topical literature from said other subcategories, each document in each said directly related topical literature including at least one unstructured field;

extracting all phrases from said unstructured field of said directly related topical literature documents;

filtering all of said extracted phrases from said directly related topical literature documents to generate a second list of high technical content phrases;

generating a co-occurrence matrix of high technical content phrases from said second list;

normalizing matrix cell values of said co-occurrence matrix to generate a normalized matrix for said unstructured field from said topical literature documents;

grouping phrases from said unstructured field of said directly related topical literature documents into thematic categories by clustering techniques on said normalized matrix;

dividing said thematic categories into a first set of categories representing specific solutions to said problem and a second set of categories that do not represent specific solutions to said problem;

generating, for each of said second set of categories, a corresponding disjoint indirectly related literature;

extracting all phrases from each said indirectly related literature;

filtering all of said phrases extracted from said indirectly related literatures to generate a second list of high technical content phrases;

grouping said high technical content phrases from said second list into thematic categories for each said indirectly related literature, the set of categories consisting of said first set of categories and said high technical content phrases from said second list into thematic categories for each said indirectly related literature to form a set of basis categories;

generating, for each of said first set of categories, and for each of said indirectly related literature thematic categories that represent potential solutions to said problems, a third list of phrases, phrase combination, and phrase co-occurrences;

filtering said third lists to remove all phrases and phrase combinations that appear in said initially retrieved documents, thereby forming filtered third lists;

determining the number of categories and the sum of frequencies over all of said basis categories for each phrase and phrase co-occurrence on said filtered third list.

ranking said phrases and phrase co-occurrences on said filtered third list by the number of categories and the sum of frequencies over all of said basis categories.

26. The method of claim 25, further comprising the steps of:

subdividing said first set of categories, or said basis categories, into thematic sub-categories; and

ranking said sub-categories for potential discovery by the number of categories in which they appear.